The Influence of Topic Difficulty, Relevance Level, and Document Ordering on Relevance Judging

Tadele T. Damessie
RMIT University
Melbourne, Australia
tadeletedla.damessie@rmit.edu.au

Falk Scholer
RMIT University
Melbourne, Australia
falk.scholer@rmit.edu.au

J. Shane Culpepper
RMIT University
Melbourne, Australia
shane.culpepper@rmit.edu.au

ABSTRACT

Judging the relevance of documents for an information need is an activity that underpins the most widely-used approach in the evaluation of information retrieval systems. In this study we investigate the relationship between how long it takes an assessor to judge document relevance, and three key factors that may influence the judging scenario: the difficulty of the search topic for which relevance is being assessed; the degree to which the documents are relevant to the search topic; and, the order in which the documents are presented for judging. Two potential confounding influences on judgment speed are differences in individual reading ability, and the length of documents that are being assessed. We therefore propose two measures to investigate the above factors: normalized processing speed (*NPS*), which adjusts the number of words that were processed per minute by taking into account differences in reading speed between judges, and normalized dwell time (*NDT*), which adjusts the duration that a judge spent reading a document relative to document length. Note that these two measures have different relationships with overall judgment speed: a direct relationship for *NPS*, and an inverse relationship for *NDT*.

The results of a small-scale user study show a statistically significant relationship between judgment speed and topic difficulty: for easier topics, assessors process more quickly (higher *NPS*), and spend less time overall (lower *NDT*). There is also a statistically significant relationship between the level of relevance of the document being assessed and overall judgment speed, with assessors taking less time for non-relevant documents. Finally, our results suggest that the presentation order of documents can also affect overall judgment speed, with assessors spending less time (smaller *NDT*) when documents are presented in *relevance order* than *docID order*. However, these ordering effects are not significant when also accounting for document length variance (*NPS*).

Keywords

user studies; dwell time; order effects; query difficulty

ADCS '16, December 05 - 07, 2016, Caulfield, VIC, Australia

ISBN 978-1-4503-4865-2/16/12...$15.00

DOI: http://dx.doi.org/10.1145/3015022.3015033

1. INTRODUCTION

Test collections are the most widely-used framework for the evaluation of the effectiveness of information retrieval (IR) systems, and consist of a representative set of search topics, a collection of documents to search, and, for each topic-document pair that is returned by a system, a relevance judgment that indicates whether the document was an appropriate response for the topic being considered. The number of relevance judgments directly affects the accuracy with which system effectiveness can be measured; however, they are expensive to obtain, since human assessors are required to determine the relevance relationship between a search topic and a document. Understanding the factors that contribute to the amount of time that it takes assessors to make relevance judgments is therefore an issue of critical importance for the evaluation of IR system.

In this work, we study the relationship between the amount of time that assessors need to make relevance judgments and three key factors that may influence this process, leading to the following research questions:

Research Question (1): *What is the relationship between the time that assessors need to make relevance judgments and the difficulty of the search topic for which such judgments are being made?*

Research Question (2): *Is there a relationship between assessor judgment time and the level of relevance of the documents being judged?*

Research Question (3): *Does the presentation order of documents have an impact on the amount of time needed to perform relevance judgments?*

A direct measure of the assessor effort required when making relevance judgments is their *dwell time*, defined as the amount of time from when an assessor is first presented with a document to judge, until they enter their judgment into an online rating system. However, using this raw quantity to investigate the research questions outlined above leads to at least two confounding factors: different assessors have different levels of reading ability, as reflected in different reading speeds; and, the documents to be judged are of different lengths. We therefore propose two measures of judgment time to account for these issues. *Normalized dwell time* (*NDT*) accounts for differences is dwell time across assessors using geometric averaging. *Normalized processing speed* (*NPS*) build on *NDT* by normalizing the reading speed of the assessor with respect to the length of each document – resulting in an average "words per minute" processing speed. Note that these two measures have different relationships with assessor judgment time: a *higher NPS* means that the assessor was *faster* on average when judging a document, while a *lower NDT* similarly shows that the assessor was *faster*.

Table 1: A breakdown comparison of previous research on topic difficulty, relevance, dwell time, presentation order, and system effects in IR user studies explored in this work.

Topic difficulty		Degree of relevance		Presentation order	Dwell time		Document length
user view	system view	binary relevance	graded relevance		for relevance assessment	for information need	
[7, 9, 17, 18]	[1, 3, 6, 7, 10, 20, 21, 27, 34, 35]	[14]	[24]	[7]	[33]	[4, 11, 12, 14, 14, 15, 16, 19, 25, 26, 28, 29, 32, 33, 33]	[16, 19, 28]

Our contributions: We analyze the association between the time taken to perform relevance judgments and topic difficulty, degree of document relevance, and presentation order. The results provide new insights into various aspects of the relevance judgment process, which is the most resource intensive component when building new test collections, and have implications when designing new human assessment approaches. We further propose and report on two measures to take into account potential confounding factors when considering judgment time, providing additional insights into the analysis and interpretation of judgment time during relevance assessment exercises.

2. RELATED WORK

We start the review section by quantifying relevance in the context of IR, followed by a review of studies on topic difficulty. Next, we review work examining dwell time in relevance assessment. Finally we review work describing the effect of document order on relevance judgments. Table 1 categorizes the relevant work into five main classes. Our study focuses mainly on the correlation between the classes as highlighted in Section 1 of the research questions.

2.1 Relevance

Relevance is a central concept in IR. Researchers have devoted a substantial portion of their research efforts to define and contextualize relevance [2, 5, 22, 23]. Saracevic [22] made a further distinction between the *system view* and the *user view* of relevance, and argue that these two views are not necessarily in agreement. Algorithms, measurements and evaluation metrics attempt to provide system measures of relevance. However, user dimensions of relevance are cognitive, affective and situational, and can encompasses topicality, utility and cognitive matching. Our understanding of relevance in this study is informed by both of the system and user views of IR.

2.2 Topic difficulty

Topic difficulty has many different guises in IR. Query performance, query hardness, query quality and query ambiguity are but a few common formulations of topic difficulty. The difficulty of a topic can be viewed either from a *system view* [1, 3, 6, 10, 20, 21, 27, 34, 35] or from a *user view* [9, 17]. Using techniques to predict the difficulty of a topic, there are two major classes of system topic difficulty [8, 27], namely *post-retrieval* and *pre-retrieval* topic difficulty. Topic difficulty not only affects systems and users. It can also play an important role in user agreement during assessment exercises [7, 24].

The work of Aslam and Pavlu [1], Carterette et al. [3], Pérez-Iglesias and Araujo [21], Shtok et al. [27], Yom-Tov et al. [34], and Cronen-Townsend and Croft [6] all belong to the search dependent post-retrieval class. The search dependent class of topic performance prediction uses relevance scores for a query, and a given retrieval model to flag a query as hard, whereas He and Ounis [10], Zhao et al. [35], and Mothe and Tanguy [20] focus on search independent pre-retrieval query performance prediction techniques. A pre-retrieval technique uses measures like specificity, ambiguity and the relationship between query terms to classify queries into easy and hard [8, 10, 35].

In order to investigate user topic difficulty, Koopman and Zuccon [17] asked assessors how difficult documents were to assess, and analyzed several cognitive factors affecting relevance judgments. The study reported that the difficulty of interpreting a query, and the presence of multiple aspects of a query can contribute to user query difficulty. Hauff et al. [9] compared user and system query performance prediction for topics given a query and its description, and concluded that assessors were able to distinguish between "good" (effective) and "bad" (ineffective). A similar user experiment was conducted by Lioma et al. [18], who gathered user opinions for queries based on personal experiences of users, and found that users underestimated system hard queries when using only topics, but their assessment showed improvement when the users were asked to assess against individual causes of topic difficulty such as ambiguity and specificity. Improvements in user accuracy when extra information is provided for difficult topics aligns with observations made by Kelly [13], who recommended using a post hoc approach when attempting to analyze dependent variables in interactive information retrieval scenarios.

2.3 Dwell time

Dwell time is another important factor in user focused relevance studies [4, 11, 12, 14, 15, 16, 19, 25, 26, 28, 29, 32, 33]. Adequately interpreting the interaction of time with topic difficulty and document ordering requires a careful analysis of the context in which the measurements are made. For example, in a typical interactive information search scenario, users may spend more time reading documents that they find relevant or "interesting" with respect to their information need. This may not be the case in other tasks such as relevance assessment where the main interaction with a document is to determine if a document is related to the information need. Hence, trying to associate longer reading time with document relevance and not considering the context of the task may lead to inappropriate conclusions.

User focused relevance studies have started incorporating time as an important variable in relevance assessment [4, 11, 12, 14, 15, 16, 19, 25, 26, 28, 29, 32, 33]. Cooper and Chen [4], Konstan et al. [16] and Seo and Zhang [25] all concluded that users spend more time reading documents that they find relevant. Konstan et al. [16] studied Usenet news group users and found a correlation between time and relevance. Though the correlation was not strong, the findings suggested that users spend longer reading news items which they find relevant than those items which are not. Cooper and

Chen [4] used a logistic regression model to predict relevance. The predictive model was developed using variables characterizing a web-based catalog search. Their findings supported the hypothesis that users spend more time on relevant sessions than non-relevant sessions.

Kelly and Cool [15] studied the relationship between familiarity and reading time. Though the finding did not observe any significance, the study concluded the general trend that subjects who had higher topic familiarity spent less time reading documents.

Kelly [12] stressed the need to exercise caution when using implicit measurements to infer relevance, and argued that the context should also be taken into account. Using dwell time to infer relevance without considering the purpose of users interaction with documents could lead to an incorrect interpretation of the results. A person reading a document when judging relevance interacts differently than a user trying to find relevant documents in order to satisfy an information need.

The distinction between users who search to identify informative documents, and those of relevance judges who search for evidence of relevance in a document was further explored by Yilmaz et al. [33]. Their study modeled a two stage process for users – initial assessment and extract utility – where users make adjustments to their expectation followed by a commitment to read an entire document. Unlike users, assessors commit to find evidence of relevance in a document throughout the entire assessment task. The study concluded that judges are likely to spend more time on documents requiring higher effort to find relevant information. High-effort documents are documents which are too long or too difficult to read, and require assessors to exert extra effort to find relevant information.

Smucker and Clarke [28] proposed a time-biased evaluation metric which takes into account the time a user takes to reach a particular document in a ranked search result list. Smucker and Jethani [29] used time as an indicator of assessor error in relevance judgments. On average assessors spend more time making inaccurate judgments than when making correct judgments.

Kelly and Belkin [14] reported a correlation between reading time and relevance. The distribution of documents identified as relevant and non-relevant by the participants were 43% (240) documents and 57% (321) documents respectively. The study found no significant reading time difference between relevant and non-relevant documents. The study did not consider the length of documents, though findings by Konstan et al. [16] and Morita and Shinoda [19] asserted that there is no significant correlation between time and document length. However, Smucker and Clarke [28] later reported a correlation between document length and dwell time in their study. In addition to considering document length, we also consider the variation in dwell time between individuals when performing judgments.

2.4 Document ordering

Damessie et al. [7] investigated the effect of document ordering during relevance assessments with respect to topic difficulty. The study found that order effects during relevance assessments can be amplified by both system and user topic difficulty. However, the study did not consider how document ordering affects time spent by assessors when performing judgments, which is one of the research questions that will be addressed in this study.

3. METHODOLOGY

To investigate assessor judgment time and the factors that may impact on this, we carried out a user study.

3.1 User study

Topics and Documents. Our study makes use of the TREC-7 and TREC-8 document collections, and the 4-level graded relevance judgments previously created by Sormunen [30]. There are 41 topics with relevance judgments in this dataset, of which 4 topics were selected. The topics `#356 el nino` (AAP $= 0.723$), `#410 schengen agreement` (AAP $= 0.643$), `#378 euro opposition` (AAP $= 0.046$), and `#448 ship losses` (AAP $= 0.024$) were selected as being representative of both hard and easy topics, as these four queries represent the two highest and two lowest *A*verage-of-*A*verage-*P*recision (AAP) scores on 110 runs from the 2004 TREC Robust Track submissions. AAP is an estimate of how difficult a topic is based on how effective an IR system is at finding relevant answers, measured by AP, for that topic.

We also considered topic difficulty from a user perspective by investigating the participant responses to the exit assessment questionnaire item "How easy was it to identify relevant documents for the search topic?". In line with previous work on topic hardness, a correlation between user and system topic difficulty [7] was found. We therefore continue to class topics #365 and #410 as *easy*, and topics #378 and #448 as *hard*.

For each topic, 30 documents were selected such that the distribution of their graded relevance levels was kept proportional to the total number of documents for each relevance grade, as found in the original Sormunen [30] relevance judgments file. For example, topic #448 has a total of 163 documents, of which 119 are non-relevant, 14 are marginally relevant, 27 are relevant and 3 are highly relevant. The proportion of non-relevant, marginally relevant, relevant and highly relevant documents that constitutes the 30 documents for the experiment are 22, 2, 5 and 1 respectively. However, an exception was made for topics #365 and #410, where only one highly relevant document exists in the relevance judgments file. For both cases, the highly relevant document was included in the list of the 30 experimental documents. This was to ensure that all relevance levels are present for each topic, and to balance the experimental setup.

Participants and Study Design. A total of 16 graduate students were recruited at RMIT University to participate in the study. They were invited to the IR search laboratory, where their task was explained to them. Since the amount of time spent making relevance judgments is the key response variable of interest, the experiment was conducted in a quiet room to minimize external distractions. On arriving at the lab, the relevance assessment task was explained to participants, based on an instruction script. The information included an indication of the expected task time (2 hours) for judging relevance in response to two search topics, and there was a 10 minute break between each topic session. The relevance criteria for rating documents as highly relevant, relevant, marginally relevant or non-relevant [30] was provided on paper, and placed on the desk next to each assessor so that it could be referred to during the experiment. The definitions were also displayed on-screen as part of an initial practice task, to familiarize participants with the judgment interface. Participants were compensated for their time with a $30 shopping voucher.

Each subject was assigned a unique assessment identifier. A pre-assessment questionnaire was used to collect their prior level of familiarity of the query, the perceived clarity of the query descriptions and narratives, and their level of confidence in identifying relevant documents. At the end of the experiment, a post-assessment questionnaire was also used to collect their perceptions on familiarity, clarity and ease of identifying relevant documents for the search

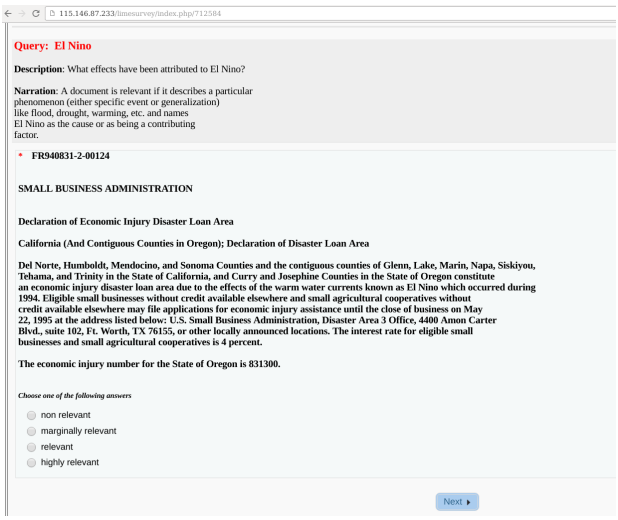

Figure 1: Relevance assessment interface developed for our assessment exercise. For each topic – document pair, the query, description, and topic narrative are presented using the original TREC information, followed the document. At the bottom of each document, the user enters their judgment. For each document judged, the relevance and dwell time (in seconds) is recorded.

topics. A five point Likert scale was used for both the pre- and post-assessment questionnaires.

The main part of the study consisted of making relevance judgments. For this, an experimental judgment interface was created. The participants were presented with a sequence of topic-document pairs, one at a time. At the top of the screen, the Title, Description and Narrative of the TREC search topic were displayed, followed by a single document. Below this was a form consisting of four radio buttons, allowing assessors to judge the document as being one of: not relevant; marginally relevant; relevant; or highly relevant. A screen-shot of the interface is provided in Figure 1. Each assessor was asked to judge two of the four topics, one easy and one hard, using the relevance assessment system. For each topic, a sequence of 30 documents had to be judged. The system recorded each response, as well as the wall clock time spent judging each document, measured from when the current document's page loaded until the time that the assessor clicked the submit button to record their relevance judgment. In addition, the presentation order of documents was controlled using decreasing relevance order (*relevance order*) and the TREC assigned identifier (*docID order*). Both orderings were rotated between assessors in order to evenly distribute the order presentation effect across assessors. Assessors were not allowed to go back and change their relevance score once they submitted a judgment.

3.2 Data analysis

Significance testing. In this work, we are concerned with the time taken to make relevance judgments during the assessment exercise. Each assessor judged two of the four topics; there are therefore a total of 960 data points (16 assessors $\times$ 2 topics $\times$ 30 documents).

To answer the first and third research questions, we analyze the response data using an unpaired Wilcoxon signed rank test; this is a non-parametric test of the null hypothesis that there is no difference in the median scores of two samples. The second research question involves the analysis of four topics (#365, #378, #410 and #448)

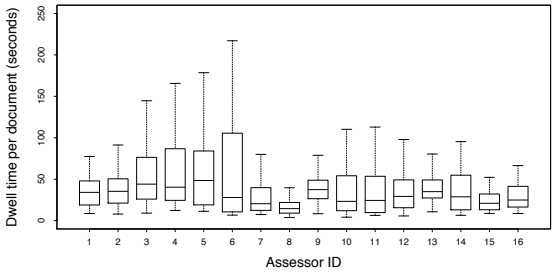

Figure 2: Raw dwell time for each individual assessor.

and four levels of relevance; a one-way ANOVA is used to test for statistical significance between groups. Where significance is detected, ANOVA is followed by a post hoc Tukey's HSD test to examine which specific pairs of items have a significant difference. For all tests, a threshold of $p < 0.05$ is taken as being indicative of statistical significance.

Normalization. This study is concerned with the time taken to make relevance judgments as an independent variable, and considers the relationship with topic difficulty, document relevance level, and presentation order as dependent variables. A raw measurement for the independent variable is therefore the wall clock time that each judge took to assess each topic-document pair. However, at least two possible confounding variables are present. First, the length of documents varies, where longer documents may simply take longer to read than shorter documents. In our study, the length of documents varied from a minimum of 118 words to a maximum of 96,910 words, with a mean and median word count of 4,695 and 666, respectively. A chi-square test shows that document lengths vary significantly across the collection (χ^2 (959) = 32,945,000, $p < 0.0001$). A second potential problem may arise due to individual differences between participants, whereby some people are faster readers than others. Figure 2 shows the distribution of raw times taken by each assessor when judging 60 documents. An ANOVA shows significant differences in mean dwell time across the 16 participants (F(15, 944) = 5.547, $p < 0.0001$).

Given the significant effects of individual judgment speed and document length, we propose two different normalizations of the raw clock time taken to make relevance judgments: *normalized dwell time* (NDT) and *normalized processing speed* (NPS). The former accounts for the differences in speed per assessor, and the latter accounts for different document lengths.

To normalize the time taken to judge the relevance of individual topic-document pairs and obtain normalized dwell time (*NDT*), we use geometric averaging [31]. First, $\log(\textit{time})$ is calculated for all raw scores. Next, the mean of these $\log(\textit{time})$ scores is calculated for each assessor (μ_a). In addition, a global mean (μ) is calculated based on all of the transformed data points. Each of the individual transformed time data points is then adjusted, by adding on the global mean (μ), and subtracting off the per-assessor mean (μ_a). Finally, each score is reverted to the original scale by applying the antilog:

$$NDT = \exp^{(\log(\textit{time})+\mu-\mu_a)} \quad (1)$$

To calculate normalized processing speed (*NPS*), the previously obtained normalized dwell time (*NDT*) can be computed with respect to document length *(docLen)* counted as number of words:

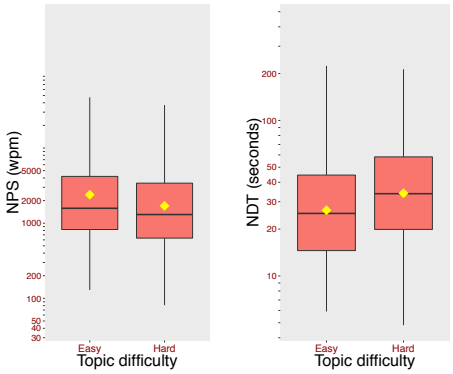

Figure 3: Normalized processing speed, *NPS* (left), and normalized dwell time, *NDT* (right), for assessors with respect to topic difficulty. Note that *NPS* and *NDT* are on a log scale.

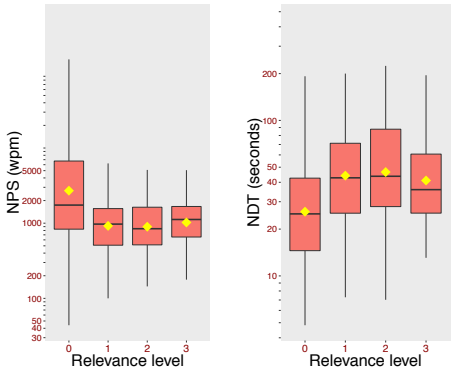

Figure 4: Normalized processing speed, *NPS* (left), and normalized dwell time, *NDT* (right), for assessors with respect to degrees of document relevance. Note that *NPS* and *NDT* are on a log scale.

$$NPS = \frac{docLen}{NDT} \tag{2}$$

It is important to note that these two measures have different relationships with the raw time scores. For *NDT*, which is based on total dwell time per document, smaller values represent *less* time in making judgments (assessors were faster), while for *NPS*, which is based on "words per minute" (wpm), smaller values equate to a *longer* time (assessors were slower). Both metrics are therefore related to more traditional work on dwell time, but care should be taken when looking at the comparisons.

4. RESULTS AND DISCUSSION

Judgment time and topic difficulty. The first research question focuses on the relationship between the time that assessors need to make relevance judgments and the difficulty of the search topic for which the judgments are being made. Figure 3 plots the distribution of times by topic difficulty (easy and hard), for both *NDT* and *NPS*. It can be seen that there is a noticeable difference in *NPS* between the easy and difficult topics. An unpaired Wilcoxon signed rank test reveals a significant difference of 265.2 in *NPS* ($p = 0.0009$; 99% confidence interval [57.3637, 484.5044]). Our judges read documents more quickly for easy topics than the hard ones. Recall that our topic sets were split for difficulty based on both system-centric (AAP) and user-centric (direct participant response about perception of difficulty) notions, and that these were directly correlated – the finding therefore applies to both user and system notions of topic difficulty. For *NDT*, which also takes individual differences between participants into account, the difference of -6.7 is also significant ($p < 0.0001$; 99% confidence interval $[-10.3146, -3.3324]$). That is, normalized dwell time (*NDT*) is significantly less for easy topics than hard topics for all assessors, which is consistent with the *NPS* comparison. In response to RQ1, the evidence suggests that assessors spend less time judging documents for easy search topics than they do for hard topics.

Judgment time and degree of relevance. The second investigated factor was how the relevance level of a document being judged might influence the time that assessors need to make their judgments. Figure 4 shows the distribution of judgment times, for *NDT* and *NRS*, split by the Sormunen relevance level of the document being judged (0–not relevant; 1–marginally relevant; 2–relevant; 3–highly relevant). A one-way ANOVA to compare the effects of the degree of relevance on processing speed for the aggregated data shows significant differences for both *NDT* (F(3, 956) = 8.152, $p < 0.0001$) and *NPS* (F(3, 956) =18.76, p < 0.0001). A Tukey's HSD post hoc test for differences between the individual levels is shown in Table 2 (see row "All").

The impact of relevance level can also be examined at other levels of aggregation, such as at the topic difficulty level, easy versus hard topics, or at the level of each topic individually. The distribution of *NPS* scores are shown for each topic in Figure 5. It can be seen that the trend of non-relevant documents leading to faster judgment times holds across all four topics.

Detailed statistical test results at different levels of granularity are presented in Table 2. We first consider the trends in *NPS*, where the analysis shows that the overall difference in processing speed is significant between any pairwise combination of non-relevant and the other levels (marginally relevant, relevant and highly relevant). The ANOVA further indicates significance when aggregating the relevance levels for easy topics, but not when aggregating the hard topics. The pairwise comparison for each degree of relevance in the easy topics shows that the difference for *NPS* is significant between marginally relevant and non-relevant documents. For difficult topics, there was no statistically significant difference in the degrees of relevance for *NPS* in our study. See Table 2 for a detailed breakdown of the pairwise significance effects.

When considering *NDT*, significant differences are shown between any pairwise combination of non-relevant and the other levels (marginally relevant, relevant and highly relevant), similar to what was observed for the *"All"* level for *NPS*. For easy topics, a significant difference in *NDT* is shown between the pairwise combinations of marginally relevant and (relevant, non-relevant). For difficult topics, pairwise significance is observed between marginally relevant and non-relevant documents. See the right-hand side of Table 2 for the detailed results of the ANOVA and pairwise significance effects for *NPS*.

To answer RQ2, the overall results suggest that assessors process non-relevant documents more quickly than marginally relevant, relevant or highly relevant documents. In other words, assessors spend less time on non-relevant documents.

Judgment time and order of presentation. The third factor that was investigated was presentation order. Here, we experimented with two common presentation orderings, namely decreasing ex-

Table 2: ANOVA and Tukey's HSD test results for the effect of different relevance levels of documents being judged.

Topic	Normalized Processing Speed (*NPS*)		Normalized Dwell Time (*NDT*)	
	ANOVA p-value	Tukey's HSD Pairwise p-value	ANOVA p-value	Tukey's HSD Pairwise p-value
All (#365,#378,#410,#448)	(F(3, 956) =8.152, $p < 0.0001$)	1-0 = 0.0023* 2-0 = 0.0158* 3-0 = 0.0191* 2-1 = 0.9999 3-1 = 0.9999 3-2 = 0.9999	(F(3, 956) =18.76, $p < 0.0001$)	1-0 < 0.0001* 2-0 < 0.0001* 3-0 = 0.0041* 2-1 = 0.9999 3-1 = 0.7037 3-2 = 0.7277
Easy (#365,#410)	(F(3, 476) =6.393, $p = 0.0003$)	1-0 = 0.0034* 2-0 = 0.0572 3-0 = 0.0767 2-1 = 0.9999 3-1 = 0.9999 3-2 = 0.9999	(F(3, 476) =7.825, $p < .0001$)	1-0 = 0.0034* 2-0 = 0.0038* 3-0 = 0.1078 2-1 = 0.8822 3-1 = 0.9973 3-2 = 0.8547
Hard (#378,#448)	(F(3, 476) =2.363, $p = 0.0705$)	– – – – – –	(F(3, 476) =13.29, $p < .0001$)	1-0 < 0.0001* 2-0 = 0.0035* 3-0 = 0.0638 2-1 = 0.6016 3-1 = 0.1981 3-2 = 0.8921
#365 (el nino)	(F(3, 236) =1.807, $p = 0.146$)	– – – – – –	(F(3, 236) =2.053, $p = 0.107$)	– – – – – –
#410 (schengen agreement)	(F(3, 236) =5.358, $p = 0.0014$)	1-0 = 0.0184* 2-0 = 0.0417* 3-0 = 0.0953 2-1 = 0.9999 3-1 = 0.9998 3-2 = 0.9999	(F(3, 236) =4.653, $p = 0.0035$)	1-0 = 0.0539 2-0 = 0.0408* 3-0 < 0.1338 2-1 = 0.9919 3-1 = 0.9996 3-2 = 0.9986
#378 (euro opposition)	(F(3, 236) =2.595, $p = 0.0532$)	– – – – – –	(F(3, 236) =5.132, $p = 0.0019$)	1-0 = 0.0077* 2-0 = 0.1002 3-0 = 0.1644 2-1 = 0.9384 3-1 = 0.9109 3-2 = 0.9996
#448 (ship losses)	(F(3, 236) =1.647, $p = 0.179$)	– – – – – –	(F(3, 236) =7.92, $p < 0.0001$)	1-0 < 0.0001* 2-0 = 0.1361 3-0 = 0.8602 2-1 = 0.3997 3-1 < 0.0271 3-2 = 0.6754

pected relevance order (*relevance order*) and document identifier order (*docID order*). In *relevance order*, documents are ordered from highest to lowest document relevance based on the Sormunen ground truth judgments. In *docID order*, documents are simply sorted using their document identifier and presented to assessors. The distribution of *NDT* and *NPS* times are shown, by document order, in Figure 6. It can be seen that judgment speed was slightly faster, on average, when documents are shown to assessors in relevance order. An unpaired Wilcoxon signed rank test shows that these differences are significant for *NDT* ($p = 0.0020$, with a 99% confidence interval of $[-7.3724, -0.6642]$, and a difference of -3.96 in assessors' *NDT* between *relevance order* and *docID order*), but not for *NPS* ($p = 0.0660$, with a a 99% confidence interval of $[-60.9352, 380.4462]$ and a *NPS* difference of 147.9 words per minute). This difference might be due to the placement of documents of different lengths in the two ordering approaches, with document length only being taken into account in *NPS* but not *NDT*. We plan to investigate this effect further in future work.

Overall, for RQ3, the results suggest that assessors spend less time when documents are presented in relevance order, but the effect is weak, and not significant with respect to *NPS*.

A final view into our results are representative *NPS* and *NDT* per-document breakdowns for the hard query `euro opposition`. Figure 7 shows the effect of presentation ordering for our two dwell time normalizations. This simple perspective reinforces the inverse nature of our two metrics, and provides an interesting view into the behaviour of the assessors during the exercise when faced with different orderings and degrees of relevance. Many interesting research opportunities still remain as we seek to better understand the complex interplay between topic variability and human behaviour during relevance assessments.

5. CONCLUSION

This study investigated the time that assessors need to make relevance judgments, and the influence of three factors: topic diffi-

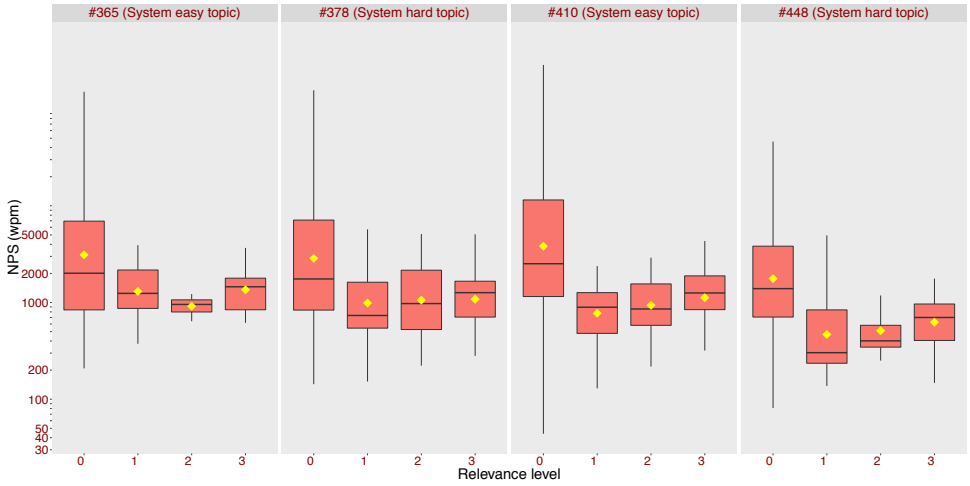

Figure 5: Normalized processing speed (*NPS*) for each level of relevance, per topic, with the mean shown as a yellow diamond. Note that *NPS* is on a log scale.

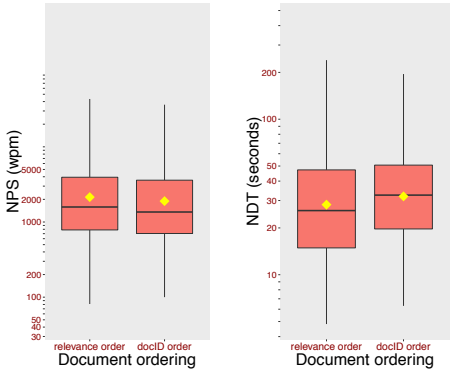

Figure 6: Normalized processing speed, *NPS* (left), and normalized dwell time, *NDT* (right), for assessors with respect to document order during assessment. Note that *NPS* and *NDT* are in log scale.

culty, relevance of the underlying document, and document ordering. Two measures, normalizing for individual speed differences (*NDT*) and for document length (*NPS*) were proposed and reported.

Our first research question examined the correlation between judgment time and topic difficulty in relevance assessment. Our findings indicate a significant relationship between dwell time and topic difficulty for both *NPS* and *NDT*: assessors are faster when judging documents of easy topics (processing more words per minute, and spending less dwell time on such documents).

The second research question investigated the relationship between judgment time and the relevance level of the document being judged. Our analysis demonstrated a significant effect on dwell time based on the level of document relevance. The pairwise overall post hoc analysis using Tukey's HSD showed that there is a difference between any pairing of non-relevant and marginally relevant, relevant, or highly relevant documents for *NDT* and *NPS*. At the level of easy versus hard topics, an effect was observed in easy topics between marginally relevant and non-relevant pairings for *NPS*; for *NDT*, an effect was also observed between marginally relevant / non-relevant, and relevant / non-relevant pairs. This suggests that assessors can more easily identify non-relevant documents, and spend less time on these than on documents of other relevance levels.

The third research question focused on the impact of document presentation ordering and assessor judgment time. The results showed a significant effect for *NDT*: assessors spend less time when documents are presented in *relevance order* than in *docID order*. However, the results for *NPS* were inconclusive. In future work, we will continue our exploration of the topic effects, and the role that system and user hardness play in query difficulty.

Acknowledgment This work was supported by the Australian Research Council's *Discovery Projects* Scheme (DP140102655 and DP140103256). Shane Culpepper is the recipient of an Australian Research Council DECRA Research Fellowship (DE140100275).

References

[1] J. A. Aslam and V. Pavlu. Query hardness estimation using jensen-shannon divergence among multiple scoring functions. In *Proc. ECIR*, pages 198–209, 2007.

[2] P. Borlund. The concept of relevance in IR. *J. Amer. Soc. Inf. Sc. Tech.*, 54(10):913–925, 2003.

[3] B. Carterette, V. Pavlu, H. Fang, and E. Kanoulas. Million query track 2009 overview. In *Proc. TREC*, 2009.

[4] M. D. Cooper and H.-M. Chen. Predicting the relevance of a library catalog search. *J. Amer. Soc. Inf. Sc. Tech.*, 52(10):813–827, 2001.

[5] W. S. Cooper. A definition of relevance for information retrieval. *Inf. Stor. and Ret.*, 7(1):19–37, 1971.

[6] S. Cronen-Townsend and W. B. Croft. Quantifying query ambiguity. In *Proc. HLT*, pages 104–109, 2002.

[7] T. T. Damessie, F. Scholer, K. Järvelin, and J. S. Culpepper. The effect of document order and topic difficulty on assessor agreement. In *Proc. ICTIR*, pages 73–76, 2016.

[8] C. Hauff, D. Hiemstra, and F. de Jong. A survey of pre-retrieval query performance predictors. In *Proc. CIKM*, pages 1419–1420, 2008.

[9] C. Hauff, D. Kelly, and L. Azzopardi. A comparison of user and system query performance predictions. In *Proc. CIKM*, pages 979–988, 2010.

[10] B. He and I. Ounis. Inferring query performance using pre-retrieval predictors. In *Proc. SPIRE*, pages 43–54, 2004.

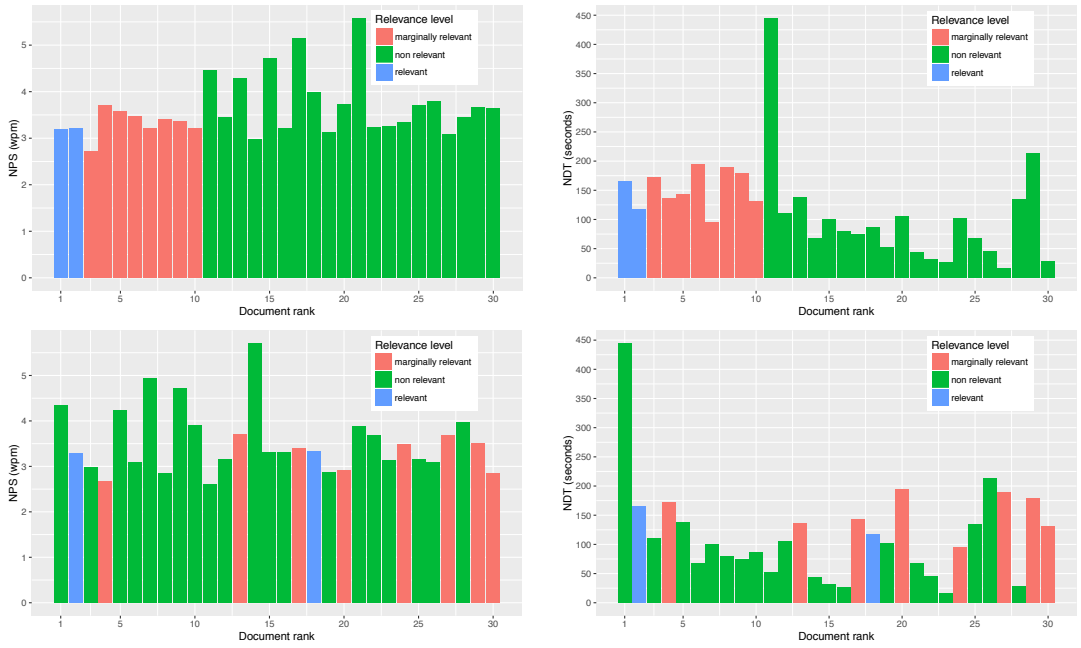

Figure 7: Dwell time and processing speed for the hard topic `euro opposition`. The top two graphs show NPS and NDT when documents are presented in relevance order, and the bottom two graphs show the results when presented in TREC ID order. These representations clearly show the expected inversion between NPS and NDT with respect to the assessors.

[11] M. Kellar, C. Watters, J. Duffy, and M. Shepherd. Effect of task on time spent reading as an implicit measure of interest. *J. Amer. Soc. Inf. Sc. Tech.*, 41(1):168–175, 2004.

[12] D. Kelly. Implicit feedback: Using behavior to infer relevance. In *New Directions in Cognitive Information Retrieval*, pages 169–186. Springer, 2005.

[13] D. Kelly. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3 (1–2):1–224, 2009.

[14] D. Kelly and N. J. Belkin. Reading time, scrolling and interaction: exploring implicit sources of user preferences for relevance feedback. In *Proc. SIGIR*, pages 408–409, 2001.

[15] D. Kelly and C. Cool. The effects of topic familiarity on information search behavior. In *Proc. JCDL*, pages 74–75, 2002.

[16] J. A. Konstan, B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon, and J. Riedl. Grouplens: applying collaborative filtering to usenet news. *Comm. ACM*, 40(3):77–87, 1997.

[17] B. Koopman and G. Zuccon. Why assessing relevance in medical IR is demanding. In *Medical Information Retrieval Workshop at SIGIR 2014*, 2014.

[18] C. Lioma, B. Larsen, and H. Schutze. User perspectives on query difficulty. In *Proc. ICTIR*, pages 3–14, 2011.

[19] M. Morita and Y. Shinoda. Information filtering based on user behavior analysis and best match text retrieval. In *Proc. SIGIR*, pages 272–281, 1994.

[20] J. Mothe and L. Tanguy. Linguistic features to predict query difficulty. In *SIGIR Workshop on Predicting Query Difficulty: Methods and Applications*, pages 7–10, 2005.

[21] J. Pérez-Iglesias and L. Araujo. Standard deviation as a query hardness estimator. In *Proc. SPIRE*, pages 207–212, 2010.

[22] T. Saracevic. Relevance: A review of the literature and a framework for thinking on the notion in information science. part II: nature and manifestations of relevance. *J. Amer. Soc. Inf. Sc. Tech.*, 58(13):1915–1933, 2007.

[23] T. Saracevic. Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance. *J. Amer. Soc. Inf. Sc. Tech.*, 58(13):2126–2144, 2007.

[24] F. Scholer, D. Kelly, W.-C. Wu, H. S. Lee, and W. Webber. The effect of threshold priming and need for cognition on relevance calibration and assessment. In *Proc. SIGIR*, pages 623–632, 2013.

[25] Y.-W. Seo and B.-T. Zhang. A reinforcement learning agent for personalized information filtering. In *Proc. IUI*, pages 248–251, 2000.

[26] M. Shokouhi, R. White, and E. Yilmaz. Anchoring and adjustment in relevance estimation. In *Proc. SIGIR*, pages 963–966, 2015.

[27] A. Shtok, O. Kurland, D. Carmel, F. Raiber, and G. Markovits. Predicting query performance by query-drift estimation. *ACM Trans. Information Systems*, 30(2):11, 2012.

[28] M. D. Smucker and C. L. Clarke. Time-based calibration of effectiveness measures. In *Proc. SIGIR*, pages 95–104, 2012.

[29] M. D. Smucker and C. P. Jethani. Time to judge relevance as an indicator of assessor error. In *Proc. SIGIR*, pages 1153–1154, 2012.

[30] E. Sormunen. Liberal relevance criteria of trec – Counting on negligible documents? In *Proc. SIGIR*, pages 324–330, 2002.

[31] A. Turpin, F. Scholer, S. Mizzaro, and E. Maddalena. The benefits of magnitude estimation relevance assessments for information retrieval evaluation. In *Proc. SIGIR*, pages 565–574, 2015.

[32] R. Villa and M. Halvey. Is relevance hard work? Evaluating the effort of making relevant assessments. In *Proc. SIGIR*, pages 765–768, 2013.

[33] E. Yilmaz, M. Verma, N. Craswell, F. Radlinski, and P. Bailey. Relevance and effort: an analysis of document utility. In *Proc. CIKM*, pages 91–100, 2014.

[34] E. Yom-Tov, S. Fine, D. Carmel, and A. Darlow. Learning to estimate query difficulty: Including applications to missing content detection and distributed information retrieval. In *Proc. SIGIR*, pages 512–519, 2005.

[35] Y. Zhao, F. Scholer, and Y. Tsegay. Effective pre-retrieval query performance prediction using similarity and variability evidence. In *Proc. AIRS*, pages 52–64. 2008.